

Asymptotic regimes for the occupancy scheme of multiplicative cascades

Jean Bertoin

*Laboratoire de Probabilités, Université Pierre et Marie Curie
and DMA, Ecole Normale Supérieure
175, rue du Chevaleret
F-75013 Paris, France*

Summary. In the classical occupancy scheme, one considers a fixed discrete probability measure $\mathbf{p} = (p_i : i \in \mathcal{I})$ and throws balls independently at random in boxes labeled by \mathcal{I} , such that p_i is the probability that a given ball falls into the box i . In this work, we are interested in asymptotic regimes of this scheme in the situation induced by a refining sequence $(\mathbf{p}(k) : k \in \mathbb{N})$ of random probability measures which arise from some multiplicative cascade. Our motivation comes from the study of the asymptotic behavior of certain fragmentation chains.

Key words. Occupancy scheme, multiplicative cascade, asymptotic regime, homogeneous fragmentation.

A.M.S. Classification. 60 F 15, 60 J 80.

e-mail. jbe@ccr.jussieu.fr

1 Introduction

The occupancy scheme is a simple urn model in probability theory that possesses a variety of applications to statistics, combinatorics, and computer science. These include, for instance, species sampling [7, 12], analysis of algorithms [9], learning theory [6], etc. The books by Johnson and Kotz [15] and by Kolchin *et al.* [18] are standard references.

This model is often depicted as balls-in-bins. Typically, we denote by $\text{Prob}_{\mathcal{I}}$ the space of probability measures on some countable set of indices \mathcal{I} , so each $\mathbf{p} \in \text{Prob}_{\mathcal{I}}$ can be identified as a family $\mathbf{p} = (p_i : i \in \mathcal{I})$ of nonnegative real numbers with $\sum_{i \in \mathcal{I}} p_i = 1$. Given some $\mathbf{p} \in \text{Prob}_{\mathcal{I}}$, one throws balls successively and independently in a fixed series of boxes labeled by indices i in \mathcal{I} , and assumes that each ball has probability p_i of falling into the box i . For every integers j, n with $j \leq n$, we denote by $N_{n,j}^{\mathbf{p}}$ the number of boxes containing exactly j balls when n balls

have been thrown, and by

$$N_n^{\mathbf{P}} := \sum_{j=1}^{\infty} N_{n,j}^{\mathbf{P}}$$

the total number of occupied boxes.

We consider here a variant of this occupancy scheme which corresponds to a nested family of boxes. This is conveniently described in terms of the genealogical structure of populations, so we start by recalling some notions in this area. We introduce the infinite *genealogical tree*

$$\mathcal{T} := \bigcup_{k=0}^{\infty} \mathbb{N}^k,$$

with $\mathbb{N} := \{1, 2, \dots\}$ and the convention $\mathbb{N}^0 := \{\emptyset\}$. The elements of \mathcal{T} are called *individuals*, and for every integer k , the k -th generation of \mathcal{T} is formed by the individuals in \mathbb{N}^k . The boundary $\partial\mathcal{T} = \mathbb{N}^{\mathbb{N}}$ of \mathcal{T} is the set of infinite sequences $\ell = (\ell_1, \ell_2, \dots)$ of positive integers, which we call *leaves*. For each leaf $\ell = (\ell_1, \ell_2, \dots)$ and each integer k , we write $\ell^{(k)} = (\ell_1, \dots, \ell_k)$ for the ancestor of ℓ at generation k . Conversely, for every individual at generation k , say $i \in \mathbb{N}^k$, we denote by $\partial\mathcal{T}_i$ the subset of leaves whose ancestor at generation k is i . In particular, the root \emptyset of the genealogical tree should be viewed as the progenitor of the entire population, and $\partial\mathcal{T}_{\emptyset} = \partial\mathcal{T}$.

Then consider some probability measure, say \mathbf{P} , on $\partial\mathcal{T}$, and imagine that we sample a sequence $\lambda_{(1)}, \lambda_{(2)}, \dots$ of i.i.d. random leaves according to the law \mathbf{P} . For every fixed integers $k, n \in \mathbb{N}$, we denote by $N_n^{(k)}$ the number of ancestors at generation k of the first n leaves :

$$\begin{aligned} N_n^{(k)} &:= \text{Card} \left\{ \lambda_{(m)}^{(k)} : m \leq n \right\} \\ &= \text{Card} \left\{ i \in \mathbb{N}^k : \text{Card} \left(\partial\mathcal{T}_i \cap \{ \lambda_{(1)}, \dots, \lambda_{(n)} \} \right) \geq 1 \right\}. \end{aligned}$$

More precisely, we may also consider for every integer $1 \leq j \leq n$

$$N_{n,j}^{(k)} := \text{Card} \left\{ i \in \mathbb{N}^k : \text{Card} \left(\partial\mathcal{T}_i \cap \{ \lambda_{(1)}, \dots, \lambda_{(n)} \} \right) = j \right\},$$

the number of individuals i at generation k such that the boundary $\partial\mathcal{T}_i$ of the subtree that stems from i contains exactly j leaves among $\{ \lambda_{(1)}, \dots, \lambda_{(n)} \}$. The connexion with the classical occupancy scheme may be better understood by viewing the random leaves as balls which are thrown on the boundary of the tree, and then imagining that each ball falls down following the branch from the leaf to the root \emptyset . Each individual i can be thought of as a box, and if balls are thrown randomly according to the probability measure \mathbf{P} on the boundary of the tree, then the probability that some given ball passes through the box i at generation k is

$$p_i(k) := \mathbf{P}(\partial\mathcal{T}_i).$$

Clearly, $\mathbf{p}(k) := (p_i(k) : i \in \mathbb{N}^k)$ defines a probability measure on \mathbb{N}^k for each generation k , and the sequence of discrete probability measures $(\mathbf{p}(k) : k \in \mathbb{N})$ determines \mathbf{P} .

Recently, Gneden *et al.* [11] have considered asymptotic laws for a randomized version of the classical occupation scheme, where the discrete probability $\mathbf{p} \in \text{Prob}_{\mathcal{T}}$ is random (more

precisely, \mathbf{p} is obtained from the atoms of some Poisson point measure on $]0, \infty[$. In the present work, we will be interested in a situation when the probability measure \mathbf{P} on $\partial\mathcal{T}$ (and therefore also each probability measure $\mathbf{p}(k)$ on \mathbb{N}^k) is random. So henceforth, *conditionally on \mathbf{P}* , each leaf $\lambda_{(1)}, \dots$ is picked randomly according to \mathbf{P} and independently of the others. This is equivalent to assuming that the sequence of random leaves is exchangeable with de Finetti measure \mathbf{P} . More precisely, we shall assume that \mathbf{P} is given by some *multiplicative cascade*; see Liu [19] and the references therein. This means that we consider first some random probability measure $\boldsymbol{\varrho} = (\varrho_1, \dots)$ in $\text{Prob}_{\mathbb{N}}$ and assign to each individual $i = (i_1, \dots, i_k)$ of the genealogical tree an independent copy $\boldsymbol{\varrho}(i)$ of $\boldsymbol{\varrho}$. Roughly speaking, $\boldsymbol{\varrho}(i)$ describes how the mass $p_i(k) = \mathbf{P}(\partial\mathcal{T}_i)$ is splitted to the subsets of leaves $\partial\mathcal{T}_{ij}$ for $j \in \mathbb{N}$, where $ij = (i_1, \dots, i_k, j)$ denotes the j -th child of the individual i at generation $k+1$. Specifically, $\varrho_j(i)$ is the portion of the mass of i inherited by the child ij , i.e.

$$\frac{p_{ij}(k+1)}{p_i(k)} = \varrho_j(i),$$

so that, by iteration,

$$p_i(k) = \varrho_{i_1}(\emptyset) \times \varrho_{i_2}(i^{(1)}) \times \dots \times \varrho_{i_k}(i^{(k-1)}) \quad (1)$$

where $i^{(k')} = (i_1, \dots, i_{k'})$ denotes the ancestor of i at generation $k' \leq k$. Clearly, for each integer k , $\mathbf{p}(k) = (p_i(k) : i \in \mathbb{N}^k)$ now defines a random probability measure on \mathbb{N}^k , and we can identify the conditional laws

$$\mathcal{L}(N_n^{(k)} \mid \mathbf{p}(k)) = \mathcal{L}(N_n^{\mathbf{p}(k)}) \quad \text{and} \quad \mathcal{L}(N_{n,j}^{(k)} \mid \mathbf{p}(k)) = \mathcal{L}(N_{n,j}^{\mathbf{p}(k)}). \quad (2)$$

Our main purpose is to determine the asymptotic regimes of the numbers of occupied boxes $N_{n,j}^{(k)}$ and $N_n^{(k)}$ when both n and k tend to infinity. It is easily seen from routine estimates that non-degenerate limits should occur when $k \approx \ln n$. Since both k and n are integers, a natural regime thus could be $k = \lfloor a \ln n \rfloor$ for some real number $a > 0$, where the notation $\lfloor \cdot \rfloor$ refers to the integer part. It turns out that this is actually too crude, in the sense that for $k = \lfloor a \ln n \rfloor$, the asymptotic behavior of $N_{n,j}^{(k)}$ does not only depend on a , but also on the oscillations of the fractional part $\{a \ln n\}$. Indeed, we shall establish a law of large numbers for $N_{n,j}^{(k)}$ and a central limit theorem for $N_n^{(k)}$ when $k, n \rightarrow \infty$ in such a way that $k = a \ln n + b + o(1)$ for fixed real numbers a and b in certain intervals.

Our approach essentially combines uniform probability estimates for the classical occupancy scheme and information about asymptotic behaviors in multiplicative cascades which can be gleaned from the literature and will be reviewed in Section 2. In particular, the analysis of multiplicative cascades relies crucially on the natural connexion with a class of branching random walks, and more precisely, on their large deviations behaviors whose descriptions are due to Biggins [5]. The main results about the asymptotic regimes in our model will be presented and proved in Section 3. They include a law of large numbers and a central limit theorem mentioned above; we will also study asymptotics of the shattering generation, i.e. the lowest generation at which no box contains more than a fixed number of balls. Finally, we shall conclude this work by discussing some interpretations of the present results in the framework of homogeneous fragmentation processes, which provided the initial motivation for this work.

2 Preliminaries

2.1 Some limit theorems for the occupancy scheme

In this section, we lift from the literature on urn models a law of large numbers and a central limit theorem for the number of occupied boxes that will be useful in our study.

Given an arbitrary discrete probability measure $\mathbf{p} = (p_i : i \in \mathcal{I}) \in \text{Prob}_{\mathcal{I}}$, we define first for every $j \leq n$ the number $\bar{N}_{n,j}^{\mathbf{p}}$ of boxes occupied by more than j balls when n balls have been thrown, viz.

$$\bar{N}_{n,j}^{\mathbf{p}} := \sum_{\ell=j+1}^n N_{n,\ell}^{\mathbf{p}}.$$

In particular, for $j = 0$, $\bar{N}_{n,0}^{\mathbf{p}} = N_n^{\mathbf{p}}$. Introduce also for every real number $x \geq 0$

$$\begin{aligned} \bar{\mu}_j^{\mathbf{p}}(x) &:= \sum_{\ell=j+1}^{\infty} \frac{x^\ell}{\ell!} \sum_{i \in \mathcal{I}} p_i^\ell e^{-p_i x} \\ &= \sum_{i \in \mathcal{I}} \left(1 - e^{-p_i x} \left(1 + \dots + \frac{(p_i x)^j}{j!} \right) \right). \end{aligned}$$

We may now state the following law of large numbers which has its roots in Bahadur [1].

Lemma 1 *Let $\mathbf{p}(1), \mathbf{p}(2), \dots$ be a sequence of discrete probability measures, $(n_k, k \in \mathbb{N})$ a sequence of positive integers with $\lim_{k \rightarrow \infty} n_k = \infty$, and $j \in \mathbb{Z}_+$. Suppose that*

$$\sum_{k \in \mathbb{N}} \frac{1}{\bar{\mu}_j^{\mathbf{p}(k)}(n_k)} < \infty \quad (3)$$

and

$$\lim_{\alpha \rightarrow 1} \lim_{k \rightarrow \infty} \frac{\bar{\mu}_j^{\mathbf{p}(k)}(\alpha n_k)}{\bar{\mu}_j^{\mathbf{p}(k)}(n_k)} = 1.$$

Then

$$\lim_{k \rightarrow \infty} \frac{\bar{N}_{n_k,j}^{\mathbf{p}(k)}}{\bar{\mu}_j^{\mathbf{p}(k)}(n_k)} = 1 \quad a.s.$$

Although this result should belong to the folklore of limit theorems for urn models, we have not been able to find a precise reference where it is stated in this form, and thus we shall provide a proof. The argument relies on Poissonization, which is an important technique in this area; see, for instance, the surveys by Gneden *et al.* [10] or Holst [13].

Proof: We work first with a fixed probability measure $\mathbf{p} = (p_i : i \in \mathcal{I})$, but we replace the deterministic number of balls n by \mathbf{n}_x , where $\mathbf{n} = (\mathbf{n}_x, x \geq 0)$ is an independent standard Poisson process. The key effect of Poissonization is that now, for each $i \in \mathcal{I}$, the number of balls in the box i has the Poisson distribution with parameter $p_i x$, and that to different boxes correspond independent Poisson variables. As a consequence, the variable β_i which takes the

value 1 if more than j balls occupy the box i and 0 otherwise, has the Bernoulli distribution with parameter

$$\sum_{\ell=j+1}^{\infty} e^{-p_i x} \frac{(p_i x)^\ell}{\ell!},$$

and when i varies in \mathcal{I} , these Bernoulli variables are independent. Changing the typography, we write

$$\bar{N}_{x,j}^{\mathbf{p}} := \sum_{i \in \mathcal{I}} \beta_i$$

for the number of boxes occupied by more than j balls when \mathbf{n}_x balls have been thrown. By elementary properties of sums of independent Bernoulli variables, we see that

$$\mathbb{E}(\bar{N}_{x,j}^{\mathbf{p}}) = \bar{\mu}_j^{\mathbf{p}}(x) \quad (4)$$

and

$$\text{Var}(\bar{N}_{x,j}^{\mathbf{p}}) \leq \bar{\mu}_j^{\mathbf{p}}(x).$$

Thus Chebyshev's inequality ensures that

$$\mathbb{P}\left(\left|\frac{\bar{N}_{x,j}^{\mathbf{p}}}{\bar{\mu}_j^{\mathbf{p}}(x)} - 1\right| \geq \varepsilon\right) \leq \frac{\varepsilon^{-2}}{\bar{\mu}_j^{\mathbf{p}}(x)}$$

for every $\varepsilon > 0$.

Next, we replace the fixed probability measure \mathbf{p} by $\mathbf{p}(k)$ and take $x = \alpha n_k$ for some real number α close to 1. The bound above combined our assumptions enables us to apply the Borel-Cantelli lemma, and we get that

$$\lim_{k \rightarrow \infty} \frac{\bar{N}_{\alpha n_k, j}^{\mathbf{p}(k)}}{\bar{\mu}_j^{\mathbf{p}(k)}(n_k)} = \ell(\alpha) \quad \text{a.s.}, \quad (5)$$

with

$$\ell(\alpha) := \lim_{k \rightarrow \infty} \frac{\bar{\mu}_j^{\mathbf{p}(k)}(\alpha n_k)}{\bar{\mu}_j^{\mathbf{p}(k)}(n_k)}.$$

Recall that the number $\mathbf{n}_x = \mathbf{n}_{\alpha n_k}$ of balls which are thrown has the Poisson distribution with parameter αn_k . On the event $\{\mathbf{n}_{\alpha n_k} \geq n_k\}$, there is the bound

$$\bar{N}_{\alpha n_k, j}^{\mathbf{p}(k)} \geq \bar{N}_{n_k, j}^{\mathbf{p}(k)},$$

whereas on the complementary event we have

$$\bar{N}_{\alpha n_k, j}^{\mathbf{p}(k)} \leq \bar{N}_{n_k, j}^{\mathbf{p}(k)}.$$

Recall that $n_k \rightarrow \infty$. Plainly, if $\alpha > 1$, then

$$\lim_{k_0 \rightarrow \infty} \mathbb{P}(\mathbf{n}_{\alpha n_k} \geq n_k \text{ for all } k \geq k_0) = 1$$

whereas if $\alpha < 1$, then

$$\lim_{k_0 \rightarrow \infty} \mathbb{P}(\mathbf{n}_{\alpha n_k} < n_k \text{ for all } k \geq k_0) = 1.$$

This completes the proof, by using (5) and the assumption that $\lim_{\alpha \rightarrow 1} \ell(\alpha) = 1$. \square

Remark. If we replace the requirement (3) in Lemma 1 by the weaker $\lim_{k \rightarrow \infty} \bar{\mu}_j^{\mathbf{p}^{(k)}}(n_k) = \infty$, the same calculations yield the weak law of large numbers :

$$\lim_{k \rightarrow \infty} \frac{\bar{N}_{n_k, j}^{\mathbf{p}^{(k)}}}{\bar{\mu}_j^{\mathbf{p}^{(k)}}(n_k)} = 1 \quad \text{in probability.}$$

Next, we turn our attention to fluctuations for the number of occupied boxes. Following Hwang and Janson [14], we introduce for every fixed probability measure $\mathbf{p} = (p_i : i \in \mathcal{I}) \in \text{Prob}_{\mathcal{I}}$ and every $x \geq 0$

$$\mu_{\mathbf{p}}(x) := \bar{\mu}_0^{\mathbf{p}}(x) = \sum_{i \in \mathcal{I}} (1 - e^{-p_i x}) \quad (6)$$

and

$$\sigma_{\mathbf{p}}^2(x) := \sum_{i \in \mathcal{I}} e^{-p_i x} (1 - e^{-p_i x}) - x^{-1} \left(\sum_{i \in \mathcal{I}} x p_i e^{-p_i x} \right)^2. \quad (7)$$

These quantities provide uniform estimates for the mean and the variance of $N_n^{\mathbf{p}}$; specifically it is known from Theorem 2.3 in [14] that

$$|\mathbb{E}(N_n^{\mathbf{p}}) - \mu_{\mathbf{p}}(n)| \leq c \quad (8)$$

and

$$|\text{Var}(N_n^{\mathbf{p}}) - \sigma_{\mathbf{p}}^2(n)| \leq c, \quad (9)$$

where c denotes some numerical constant (which depends neither of n nor of \mathbf{p}). This makes the following central limit theorem quite intuitive (see Corollary 2.5 in [14], and also Dutko [8] and Karlin [16] for earlier versions).

Lemma 2 *Let $\mathbf{p}(1), \mathbf{p}(2), \dots$ be a sequence of discrete probability measures and $(k_n, n \in \mathbb{N})$ a sequence of positive integers such that*

$$\lim_{n \rightarrow \infty} \sigma_{\mathbf{p}(k_n)}^2(n) = \infty.$$

Then the number of occupied boxes is asymptotically normally distributed when n goes to infinity, in the sense that

$$\frac{N_n^{\mathbf{p}(k_n)} - \mu_{\mathbf{p}(k_n)}(n)}{\sigma_{\mathbf{p}(k_n)}(n)}$$

converges in distribution to a standard normal variable as $n \rightarrow \infty$.

We do not know whether a similar central limit theorem holds for the number $N_{n,j}^{\mathbf{p}}$ of boxes occupied by exactly j balls for $j \geq 1$.

2.2 Large deviations behaviors of multiplicative cascades

Recall that $\boldsymbol{\varrho}$ is a random probability measure on \mathbb{N} . We denote its law by ν , so ν is a probability measure on $\text{Prob}_{\mathbb{N}}$ that will be referred to as the *splitting law*. We shall always assume that this splitting law is not geometric¹, in the sense that there is no real number $r > 0$ such that with probability one, all the atoms of $\boldsymbol{\varrho}$ belong to $\{r^n, n \in \mathbb{Z}_+\}$. In particular, note that the degenerate case when $\boldsymbol{\varrho}$ is a Dirac point mass a.s. is henceforth excluded.

As it was explained in the Introduction, we consider a family $(\boldsymbol{\varrho}(i), i \in \mathcal{T})$ of independent copies of $\boldsymbol{\varrho}$ labeled by the individuals of the genealogical tree \mathcal{T} . The multiplicative cascade construction (1) defines a random probability measure $\mathbf{p}(k)$ on \mathbb{N}^k for every generation $k \in \mathbb{N}$. Our aim is to apply general asymptotic results for occupancy schemes such as Lemmas 1 and 2, and in this direction, we shall use fundamental large deviations behaviors for branching random walks that Biggins [5] established.

Taking logarithm of masses, we may encode the random probability measure $\mathbf{p}(k) = (p_i(k) : i \in \mathbb{N}^k)$ at generation k by the random point measure on \mathbb{R}_+

$$Z^{(k)}(\mathrm{d}y) := \sum \delta_{-\ln p_i(k)}(\mathrm{d}y),$$

where δ_z stands for the Dirac point mass at z and the sum in the right-hand side is taken over the individuals i at the k -th generation which have a positive mass. It then follows immediately from the structure of multiplicative cascade (1) that $(Z^{(k)}, k \in \mathbb{Z}_+)$ is a branching random walk, in the sense that for every integers $k, k' \geq 0$, $Z^{(k+k')}$ is obtained from $Z^{(k)}$ by replacing each atom z of $Z^{(k)}$ by a family $\{z + y, y \in \mathcal{Y}\}$, where \mathcal{Y} is distributed as the family of the atoms of $Z^{(k')}$ and distinct atoms z of $Z^{(k)}$ correspond to independent copies of \mathcal{Y} .

We now introduce analytic quantities defined in terms of the splitting law ν which will have an important role in the present study. First, we define the Laplace transform of the intensity measure of $Z^{(1)}$ by

$$\mathbf{L}(\theta) := \mathbb{E}(\langle Z^{(1)}, e^{-\theta \cdot} \rangle)$$

for $\theta > 0$; note that there are also the alternative expressions

$$\mathbf{L}(\theta) = \mathbb{E} \left(\sum_{j \in \mathbb{N}} \varrho_j^\theta \right) = \int_{\text{Prob}_{\mathbb{N}}} \left(\sum_{i \in \mathbb{N}} p_i^\theta \right) \nu(\mathrm{d}\mathbf{p}). \quad (10)$$

The function $\mathbf{L} :]0, \infty[\rightarrow]0, \infty]$ is convex decreasing with $\mathbf{L}(1) = 1$; we define

$$\theta_* := \inf \{ \theta > 0 : \mathbf{L}(\theta) < \infty \}, \quad (11)$$

so that $\mathbf{L}(\theta) < \infty$ when $\theta > \theta_*$. One readily sees from Hölder's inequality that $\ln \mathbf{L}$ is a convex function, and then that

$$\varphi(\theta) := \ln \mathbf{L}(\theta) - \theta \frac{\mathbf{L}'(\theta)}{\mathbf{L}(\theta)} \quad (12)$$

¹Working with a geometric splitting law would induce a phenomenon of periodicity which we shall not discuss here for simplicity. However results similar to those proven in this work can be established by the same techniques for geometric splitting laws.

is a function which decreases on $] \theta_*, \infty[$.

As $L(1) = 1$ and L decreases, we have $\varphi(1) = -L'(1) > 0$, and thus the set of $\theta \in] \theta_*, \infty[$ such that $\varphi(\theta) > 0$ is a non-empty open interval $] \theta_*, \theta^*[$, where $\theta^* > 1$ is defined by

$$\theta^* := \sup\{\theta > \theta_* : \varphi(\theta) > 0\}. \quad (13)$$

Remark : The critical parameter θ^* may be finite or infinite, and is finite whenever

$$\|\max_{i \in \mathbb{N}} \varrho_i\|_\infty = 1,$$

where $\varrho = (\varrho_i, i \in \mathbb{N})$ denotes a random probability measure on \mathbb{N} with law ν . Indeed, it is easily seen that

$$\lim_{\theta \rightarrow \infty} L(\theta)^{1/\theta} = \|\max_{i \in \mathbb{N}} \varrho_i\|_\infty,$$

and when the right-hand side equals 1, the function $g : \theta \rightarrow -\frac{\ln L(\theta)}{\theta}$ has thus limit 0 at infinity. Since g is non-negative on $[1, \infty[$ and $g(1) = 0$, g reaches its overall maximum at some location at, say, $\theta_{\max} \in]1, \infty[$. As $g'(\theta) = \theta^{-2}\varphi(\theta)$, we conclude that $\theta_{\max} = \theta^* < \infty$.

Following Biggins [4], we are now able to introduce for every $\theta > \theta_*$

$$W^{(k)}(\theta) := L(\theta)^{-k} \langle Z^{(k)}, e^{-\theta \cdot} \rangle = L(\theta)^{-k} \sum_{i \in \mathbb{N}^k} p_i(k)^\theta, \quad k \geq 0,$$

which form a remarkable family of martingales :

Lemma 3 *For every $\theta \in] \theta_*, \theta^*[$, the martingale $(W^{(k)}(\theta), k \in \mathbb{Z}_+)$ is bounded in $L^\gamma(\mathbb{P})$ for some $\gamma > 1$. Its terminal value*

$$W(\theta) := \lim_{k \rightarrow \infty} W^{(k)}(\theta)$$

is (strictly) positive a.s.

Proof: Jensen's inequality implies that for every probability measure $\mathbf{p} \in \text{Prob}_{\mathbb{N}}$ and every $\gamma > 1$, there is the upper-bound

$$\left(\sum_{i \in \mathbb{N}} p_i^\theta \right)^\gamma \leq \sum_{i \in \mathbb{N}} p_i^{\gamma(\theta-1)+1}.$$

For any $\theta > \theta_*$, we may chose $\gamma > 1$ sufficiently small such that $\gamma(\theta - 1) + 1 > \theta_*$, and we deduce that $\mathbb{E}(W^{(1)}(\theta)^\gamma) < \infty$.

We then observe that the function $f : \theta \rightarrow \theta^{-1} \ln L(\theta)$ has derivative $f'(\theta) = -\theta^{-2}\varphi(\theta)$. Thus this derivative is negative when $\theta \in] \theta_*, \theta^*[$, which means that f decreases in some neighborhood of θ . We may thus find $\gamma > 1$ sufficiently small such that

$$\frac{\ln L(\gamma\theta)}{\gamma\theta} < \frac{\ln L(\theta)}{\theta},$$

and hence $L(\gamma\theta) < L(\theta)^\gamma$. We can now apply Theorem 1 in Biggins [5], which completes the proof of the first part of our claim. Finally, the assertion that the terminal value $W(\theta) > 0$ a.s.

derives easily from the fact the probability that branching random walk $Z^{(k)}$ is extinguished at generation k equals 0 for every $k \in \mathbb{N}$. \square

In order to state the key technical result for this present study, we define for every $\theta > \theta_*$ the tilted random point measure

$$Z_\theta^{(k)}(\mathrm{d}y) := \frac{e^{-\theta y}}{L(\theta)^k} Z^{(k)}(\mathrm{d}y), \quad y \geq 0.$$

We also introduce the mean and the variance of the intensity measure of $Z_\theta^{(1)}$:

$$M(\theta) := -\frac{L'(\theta)}{L(\theta)} \quad \text{and} \quad v(\theta) = \frac{L''(\theta)}{L(\theta)} - \left(\frac{L'(\theta)}{L(\theta)} \right)^2. \quad (14)$$

Both $M(\theta)$ and $v(\theta)$ are positive quantities, and write g_θ for the (centered) Gaussian density with variance $v(\theta)$, i.e.

$$g_\theta(x) = \frac{1}{\sqrt{2\pi v(\theta)}} \exp\left(-\frac{x^2}{2v(\theta)}\right).$$

The next statement is a version of Theorem 4 in Biggins [5] specialized to our framework.

Lemma 4 *The following assertion holds with probability one:*

$$\lim_{k \rightarrow \infty} \left| \sqrt{k} Z_\theta^{(k)}([x + kM(\theta) - h, x + kM(\theta) + h]) - 2hW(\theta)g_\theta(x/\sqrt{k}) \right| = 0,$$

where the limit is uniform for $x \in \mathbb{R}$, $h \leq 1$ and θ in a compact subset of $] \theta_*, \theta^* [$.

Next, observe that if $f : \mathbb{R} \rightarrow \mathbb{R}_+$ is, say, a continuous function, then

$$\begin{aligned} \sum_{i \in \mathbb{N}^k} f(kM(\theta) + \ln p_i(k)) &= \int_{\mathbb{R}_+} f(kM(\theta) - y) Z^{(k)}(\mathrm{d}y) \\ &= \left(L(\theta) e^{\theta M(\theta)} \right)^k \int_{\mathbb{R}_+} f(kM(\theta) - y) e^{\theta(y - kM(\theta))} Z_\theta^{(k)}(\mathrm{d}y). \end{aligned}$$

Recall also that the rate function φ is defined by (12). We finally state the following limit theorem which will be useful to estimate the conditional mean number of occupied boxes given the multiplicative cascade.

Corollary 1 (large deviations behavior) *Pick $\theta \in] \theta_*, \theta^* [$ and let $f : \mathbb{R} \rightarrow \mathbb{R}_+$ be a continuous function. Assume that there exist $\alpha > 0$ and $\beta > \theta$ such that*

$$\lim_{y \rightarrow +\infty} y^\alpha f(y) = 0 \quad \text{and} \quad \lim_{y \rightarrow -\infty} e^{-\beta y} f(y) = 0,$$

so in particular $f \in L^1(e^{-\theta y} \mathrm{d}y)$. Let also $(c_k : k \in \mathbb{N})$ denote a sequence of real numbers which converges to some $c \in \mathbb{R}$. Then with probability one, we have

$$\lim_{k \rightarrow \infty} \sqrt{k} e^{-\varphi(\theta)k} \sum_{i \in \mathbb{N}^k} f(kM(\theta) + \ln p_i(k) + c_k) = \frac{e^{\theta c}}{\sqrt{2\pi v(\theta)}} \left(\int_{\mathbb{R}} f(y) e^{-\theta y} \mathrm{d}y \right) W(\theta).$$

Corollary 1 follows readily from Lemma 4 when f has bounded support. However, the derivation in the case when the function f has unbounded support is rather technical, even though our assumptions have been tailored for the purpose of the present work. We postpone the proof to the Appendix.

3 Asymptotic regimes

3.1 Main results

We now have all the technical ingredients for our study, we just need to introduce a few more notation. We shall consider the regime for pairs of integers (k, n) such that

$$k, n \rightarrow \infty \quad \text{and} \quad k - a \ln n \rightarrow b, \quad (15)$$

where $a > 0$ and $b \in \mathbb{R}$ are fixed. When $F(k, n)$ is some function depending on k and n , we shall write

$$\lim_{a,b} F(k, n)$$

for the limit of $F(k, n)$ when (k, n) follows the regime (15), of course provided that such a limit exists.

Recall that our basic datum is the splitting law ν on $\text{Prob}_{\mathbb{N}}$, and that its Laplace transform $L(\theta)$ is given by (10). Further important notions include the critical parameters θ_*, θ^* , the rate function φ , and the mean M and variance v functions, which have been defined in (11), (13), (12) and (14), respectively. The mean function M decreases continuously on $]\theta_*, \theta^*[$ and takes positive values. We denote the inverse bijection by

$$M^{-1} :]M_*, M^*[\rightarrow]\theta_*, \theta^*[,$$

where

$$M_* = \lim_{\theta \rightarrow \theta_*^-} M(\theta) \quad \text{and} \quad M^* = \lim_{\theta \rightarrow \theta_*^+} M(\theta).$$

One always has $]M_*, M^*[\subseteq]0, \infty[$, and the inclusion can be strict.

Example : These quantities are especially simple in the case when ν is the Poisson-Dirichlet distribution $\text{PD}(1)$. Indeed, one easily gets $L(\theta) = 1/\theta$ for $\theta > 0$, and then $\varphi(\theta) = -\ln \theta + 1$. One thus sees that $\theta_* = 0$ and $\theta^* = e$. Finally $M(\theta) = 1/\theta$ and $v(\theta) = 1/\theta^2$, so $M_* = 1/e$, $M^* = \infty$ and $M^{-1}(a) = 1/a$.

Finally recall from Lemma 3 that for every $\theta \in]\theta_*, \theta^*[$, $W(\theta)$ is a positive random variable which arises as the limit of a remarkable martingale. We are now able to specify regimes for the (strong) law of large numbers for the number of boxes occupied by exactly j balls in an occupation scheme driven by a multiplicative cascade.

Theorem 1 *Pick $a \in]1/M^*, 1/M_*[$ and $b \in \mathbb{R}$, and set $\theta = M^{-1}(1/a)$. Then for every integer $j > \theta$, the following limits*

$$\lim_{a,b} \sqrt{k} e^{-\varphi(\theta)k} \bar{N}_{n,j-1}^{(k)} = \left(\sum_{\ell=j}^{\infty} \frac{\Gamma(\ell - \theta)}{\ell!} \right) \frac{e^{-\theta b/a}}{\sqrt{2\pi v(\theta)}} W(\theta),$$

and

$$\lim_{a,b} \sqrt{k} e^{-\varphi(\theta)k} N_{n,j}^{(k)} = \frac{\Gamma(j-\theta) e^{-\theta b/a}}{j! \sqrt{2\pi v(\theta)}} W(\theta),$$

hold with probability one.

Remark : It may be interesting to recall that $\frac{1}{\sqrt{k}} e^{\varphi(\theta)k}$ is also the order of magnitude of the numbers of boxes of size approximately $e^{-k/a} \approx 1/n$ at generation k ; see Corollary 3 in [3]. We further stress that $\varphi(\theta) \leq 1/a$ and that this inequality is strict except when $\theta = 1$ (this can be checked directly from the observation that $\ln L$ is a strictly convex function).

Proof: We work conditionally on the random probabilities $\mathbf{p}(k)$ using (2). We aim at applying Lemma 1, and in this direction we fix some real number α close to 1 and observe that

$$\begin{aligned} \bar{\mu}_{j-1}^{\mathbf{p}(k)}(\alpha n) &= \sum_{i \in \mathbb{N}^k} \left(1 - e^{-p_i(k)\alpha n} \left(1 + \dots + \frac{(p_i(k)\alpha n)^{j-1}}{(j-1)!} \right) \right) \\ &= \sum_{i \in \mathbb{N}^k} f(k/a + \ln p_i(k) + c_k) \end{aligned}$$

with

$$f(x) = \left(1 - \left(1 + \dots + \frac{e^{(j-1)x}}{(j-1)!} \right) \exp\{-e^x\} \right)$$

and

$$c_k := \ln n + \ln \alpha - k/a = -b/a + \ln \alpha + o(1).$$

We can now check that the assumptions of Corollary 1 hold with $c = -b/a + \ln \alpha$. A straightforward calculation shows that

$$\int_{\mathbb{R}} f(y) e^{-\theta y} dy = \sum_{\ell=j}^{\infty} \frac{\Gamma(\ell-\theta)}{\ell!},$$

and then, invoking Corollary 1 and recalling that $\theta = M^{-1}(1/a)$, we deduce

$$\lim_{a,b} \sqrt{k} e^{-\varphi(\theta)k} \bar{\mu}_{j-1}^{\mathbf{p}(k)}(\alpha n) = \left(\sum_{\ell=j}^{\infty} \frac{\Gamma(\ell-\theta)}{\ell!} \right) \frac{e^{\theta(-b/a + \ln \alpha)}}{\sqrt{2\pi v(\theta)}} W(\theta). \quad (16)$$

Next, we fix $\eta \in [-1, 1]$ and define for every integer k

$$n_{k,\eta} := \lfloor \exp((k - b - \eta)/a) \rfloor.$$

Replacing b by $b + \eta$ and n by $n_{k,\eta}$ in (16), we get

$$\lim_{k \rightarrow \infty} \sqrt{k} e^{-\varphi(\theta)k} \bar{\mu}_{j-1}^{\mathbf{p}(k)}(\alpha n_{k,\eta}) = \left(\sum_{\ell=j}^{\infty} \frac{\Gamma(\ell-\theta)}{\ell!} \right) \frac{e^{\theta(\ln \alpha - (b+\eta)/a)}}{\sqrt{2\pi v(\theta)}} W(\theta).$$

Note that the right-hand side depends continuously on the variable α and that the requirement (3) in Lemma 1 is fulfilled, as $\varphi(\theta) > 0$. An application of this lemma gives

$$\lim_{k \rightarrow \infty} \sqrt{k} e^{-\varphi(\theta)k} \bar{N}_{n_{k,\eta},j-1}^{(k)} = \left(\sum_{\ell=j}^{\infty} \frac{\Gamma(\ell - \theta)}{\ell!} \right) \frac{e^{-\theta(b+\eta)/a}}{\sqrt{2\pi v(\theta)}} W(\theta),$$

with probability one. An argument of monotonicity, namely

$$n_{k,\eta} \leq n \leq n_{k,\eta'} \implies \bar{N}_{n_{k,\eta},j-1}^{(k)} \leq \bar{N}_{n,j-1}^{(k)} \leq \bar{N}_{n_{k,\eta'},j-1}^{(k)},$$

completes the proof of our first claim. The second follows immediately from the first, since $\bar{N}_{n,j}^{(k)} = \bar{N}_{n,j-1}^{(k)} - \bar{N}_{n,j}^{(k)}$. \square

We next turn our attention to finer asymptotic results for the total number of occupied boxes $\bar{N}_{n,0}^{(k)} = N_n^{(k)}$. If $\theta < 1$, that is if $a < 1/M(1)$, then we can take $j = 1$ in Theorem 1, and we get from the easy identity

$$\sum_{\ell=1}^{\infty} \frac{\Gamma(\ell - \theta)}{\ell!} = \int_0^{\infty} (1 - e^{-x}) x^{-\theta-1} dx = \frac{\Gamma(1 - \theta)}{\theta}$$

that

$$\lim_{a,b} \sqrt{k} e^{-\varphi(\theta)k} N_n^{(k)} = \frac{\Gamma(1 - \theta) e^{-\theta b/a}}{\theta \sqrt{2\pi v(\theta)}} W(\theta) \quad \text{a.s.} \quad (17)$$

Recall also from (8) that the conditional expectation of number of occupied boxes given the de Finetti measure, $\mathbb{E}(N_n^{(k)} \mid \mathbf{p}(k))$, can be approximated by $\mu_{\mathbf{p}(k)}(n) = \bar{\mu}_0^{\mathbf{p}(k)}(n)$, and that (16) provides an estimation of the latter. This gives

$$\lim_{a,b} \sqrt{k} e^{-\varphi(\theta)k} \mu_{\mathbf{p}(k)}(n) = \frac{e^{-\theta b/a} \Gamma(1 - \theta)}{\theta \sqrt{2\pi v(\theta)}} W(\theta). \quad (18)$$

Recall further (7) and (9) and consider the following approximation for the conditional variance

$$\sigma_{\mathbf{p}(k)}^2(n) = \sum_{i \in \mathcal{I}} e^{-p_i(k)n} (1 - e^{-p_i(k)n}) - (n)^{-1} \left(\sum_{i \in \mathcal{I}} n p_i(k) e^{-p_i(k)n} \right)^2.$$

Theorem 2 *Notation is the same as in Theorem 1. Provided that $\theta < 1$, we have that*

$$\lim_{a,b} \frac{\sigma_{\mathbf{p}(k)}^2(n)}{\mu_{\mathbf{p}(k)}(n)} = 2^\theta - 1 \quad \text{a.s.}$$

As a consequence, when (k, n) follows the regime (15),

$$\frac{N_n^{(k)} - \mu_{\mathbf{p}(k)}(n)}{\sqrt{(2^\theta - 1) \mu_{\mathbf{p}(k)}(n)}}$$

converges in distribution as to a standard normal variable.

It is interesting to note that the parameter b plays no rôle in the limits above, so by a standard argument based on extraction of sub-sequences, the results still hold under the weaker requirement that $k, n \rightarrow \infty$ such that $k = a \ln n + O(1)$. We do not know whether this can be extended to the more general regime when one merely requires that $k \sim a \ln n$. We also underline that the usefulness of Theorem 2 is limited by the fact that the centralizing term $\mu_{\mathbf{p}(k)}(n)$ in the numerator is random, and only its first order asymptotic (18) is known.

Proof: The calculations resemble that in the proof of Theorem 1. We start by observing that

$$\sigma_{\mathbf{p}(k)}^2(n) = \sum_{i \in \mathbb{N}^k} f(k/a + \ln p_i(k) + c_k) - (n)^{-1} \left(\sum_{i \in \mathbb{N}^k} g(k/a + \ln p_i(k) + c_k) \right)^2$$

with

$$f(x) = \exp\{-e^x\} (1 - \exp\{-e^x\}) \quad , \quad g(x) = e^x \exp\{-e^x\}$$

and

$$c_k := \ln n - k/a = -b/a + o(1) \, .$$

Easy calculations show that

$$\int_{\mathbb{R}} f(y) e^{-\theta y} dy = \frac{(2^\theta - 1) \Gamma(1 - \theta)}{\theta} \, ,$$

and

$$\int_{\mathbb{R}} g(y) e^{-\theta y} dy = \Gamma(1 - \theta) \, .$$

Applying Corollary 1 and recalling that $\theta = M^{-1}(1/a)$, we deduce that almost surely

$$\begin{aligned} \lim_{a,b} \sqrt{k} e^{-\varphi(\theta)k} \sigma_{\mathbf{p}(k)}^2(n) &= (2^\theta - 1) \frac{e^{-\theta b/a} \Gamma(1 - \theta)}{\theta \sqrt{2\pi V(\theta)}} W(\theta) \\ &= (2^\theta - 1) \lim_{a,b} \sqrt{k} e^{-\varphi(\theta)k} \mu_{\mathbf{p}(k)}(n) \, , \end{aligned}$$

which is our first claim. The second then derives from Lemma 2. □

Roughly speaking, the estimation (17) means that for $a < 1/M(1)$,

$$n \asymp e^{k/a} \implies N_n^{(k)} \asymp e^{\varphi(\theta)k} / \sqrt{k} \, .$$

We shall point out that this is no longer true when $a > 1/M(1)$, in other words that a phase transition occurs at the critical value $a = 1/M(1)$. In this direction, we consider the more general regime for pairs of integers (k, n) such that

$$k, n \rightarrow \infty \quad \text{and} \quad k \sim a \ln n \, , \tag{19}$$

for some fixed $a > 0$. Again, when $F(k, n)$ is some function depending on k and n , we shall write

$$\lim_a F(k, n)$$

for the limit (whenever it exists) of $F(k, n)$ when (k, n) follows the regime (19).

Proposition 1 *If $1/M(1) < a$, then*

$$\lim_a \frac{N_n^{(k)}}{n} = 1 \quad a.s.$$

Proof: By an argument similar to that in the proof of Theorem 1, it suffices to check that

$$\lim_{a,b} (\alpha n)^{-1} \mu_{\mathbf{p}(k)}(\alpha n) = 1 \quad a.s.$$

provided that α is sufficiently close to 1. The upperbound $\mu_{\mathbf{p}(k)}(\alpha n) \leq \alpha n$ is plain, so we shall focus on the lowerbound.

In this direction, integrating the obvious inequality $(1 + \varepsilon)x^\varepsilon \geq 1 - e^{-x}$ for $x \geq 0$ and $0 < \varepsilon \leq 1$ arbitrary, we see that

$$1 - e^{-x} \geq x - x^{1+\varepsilon}. \quad (20)$$

Replacing x by the random variable $\alpha n p_i(k)$ and summing over $i \in \mathbb{N}^k$, we deduce that

$$\mu_{\mathbf{p}(k)}(\alpha n) \geq \alpha n - (\alpha n)^{1+\varepsilon} \sum_{i \in \mathbb{N}^k} p_i(k)^{1+\varepsilon} = \alpha n - (\alpha n)^{1+\varepsilon} L(1 + \varepsilon)^k W^{(k)}(1 + \varepsilon).$$

As $W^{(k)}(1 + \varepsilon)$ is a positive martingale (in the variable k), we have $\sup_k W^{(k)}(1 + \varepsilon) < \infty$ a.s. On the other hand, we have

$$\begin{aligned} \ln(n^{1+\varepsilon} L(1 + \varepsilon)^k) &= (1 + \varepsilon) \ln n + k \ln L(1 + \varepsilon) \\ &= (1 + \varepsilon + a \ln L(1 + \varepsilon) + o(1)) \ln n. \end{aligned}$$

Recall that $-M$ is the derivative of $\ln L$ and that $\ln L(1) = 0$. Since $aM(1) > 1$, we can chose $\varepsilon > 0$ small enough so that $a \ln L(1 + \varepsilon) < -\varepsilon$. Then

$$(\alpha n)^{1+\varepsilon} L(1 + \varepsilon)^k = o(n),$$

and we conclude that

$$\liminf_{a,b} (\alpha n)^{-1} \mu_{\mathbf{p}(k)}(\alpha n) \geq 1 \quad a.s.$$

which completes the proof. □

For $\theta^* \leq 2$, a related argument also enables us to estimate the lowest generation at which all n balls fall into different boxes. More generally, recall that for every integer j , $\bar{N}_{n,j}^{(k)}$ denotes the number of boxes at generation k which contain more than j balls when n balls have been thrown. Note that this quantity increases with the number of balls n and decreases with the generation k . Define

$$\zeta_{n,j} := \min\{k \in \mathbb{N} : \bar{N}_{n,j}^{(k)} = 0\}.$$

Proposition 2 *If $\theta^* \leq j + 1$, then we have*

$$\lim_{n \rightarrow \infty} \frac{\zeta_{n,j}}{\ln n} = 1/M_* \quad a.s.$$

Proof: We first consider the randomized version with a total number of balls \mathbf{n}_x which has the Poisson law with parameter x . We write $\bar{N}_{x,j}^{(k)}$ for the number of boxes at generation k occupied by more than j balls, and consider its conditional expectation given the random probability measure $\mathbf{p}(k)$, which has been computed in (4). Using the assumption that $\theta^* \leq j+1$ at the second line below, we get

$$\begin{aligned} \mathbb{E}(\bar{N}_{x,j}^{(k)} \mid \mathbf{p}(k)) = \bar{\mu}_{\mathbf{p}(k),j}(x) &= \sum_{i \in \mathbb{N}^k} \left(1 - e^{-p_i(k)x} \left(1 + \dots + \frac{(xp_i(k))^j}{j!} \right) \right) \\ &\leq c(\theta^*) \sum_{i \in \mathbb{N}^k} (xp_i(k))^{\theta^*}, \end{aligned}$$

where

$$c(\theta^*) = \max_{y>0} y^{-\theta^*} \left(1 - e^{-y} \left(1 + \dots + \frac{y^j}{j!} \right) \right)$$

is some finite constant. Observe that the preceding upperbound can be expressed as

$$c(\theta^*) x^{\theta^*} L(\theta^*)^k W^{(k)}(\theta^*),$$

and recall that $W^{(k)}(\theta^*)$ is a martingale. The unconditional expectation $\mathbb{E}(\bar{N}_{x,j}^{(k)})$ can thus be bounded from above by $c(\theta^*) x^{\theta^*} L(\theta^*)^k$.

We next pick $a > 1/M_*$ and take $x = e^{k/a}$. Recall that

$$-\frac{\ln L(\theta^*)}{\theta^*} = M(\theta^*) = M_*.$$

We thus have

$$\ln(x^{\theta^*} L(\theta^*)^k) = k\theta^* \left(\frac{1}{a} - M_* \right),$$

and as a consequence

$$\sum_{k \in \mathbb{N}} \mathbb{E}(\bar{N}_{e^{k/a},j}^{(k)}) < \infty. \quad (21)$$

Then chose any $a' > a$ and recall that the Poisson process fulfills

$$\lim_{k_0 \rightarrow \infty} \mathbb{P} \left(\mathbf{n}_{e^{k/a}} > \lfloor e^{(k+1)/a'} \rfloor \text{ for all } k \geq k_0 \right) = 1.$$

We deduce from (21) and an argument of monotonicity that

$$\mathbb{P} \left(\bar{N}_{\lfloor e^{(k+1)/a'} \rfloor, j}^{(k)} = 0 \text{ for all integers } k \text{ sufficiently large} \right) = 1.$$

Observing that for every integer n and $k = \lfloor a' \ln n \rfloor$:

$$\zeta_{n,j} > k \implies \bar{N}_{n,j}^{(k)} \geq 1 \implies \bar{N}_{\lfloor e^{(k+1)/a'} \rfloor, j}^{(k)} \geq 1,$$

and therefore

$$\limsup_{n \rightarrow \infty} \frac{\zeta_{n,j}}{\ln n} \leq a' \quad \text{a.s.}$$

As a' can be chosen arbitrarily close to $1/M_*$, we conclude that

$$\limsup_{n \rightarrow \infty} \frac{\zeta_{n,j}}{\ln n} \leq 1/M_* \quad \text{a.s.}$$

Finally, the converse bound

$$\liminf_{n \rightarrow \infty} \frac{\zeta_{n,j}}{\ln n} \geq 1/M_* \quad \text{a.s.}$$

derives easily from Theorem 1. □

3.2 Interpretation in terms of homogeneous fragmentations

The initial motivation for this work was to gain insight on certain asymptotic regimes for homogeneous fragmentations. Roughly, the latter form coherent families of natural Markov processes with values in the space of partitions of finite sets, such that these random partitions are refining as time passes. They are closely related to multiplicative cascades of random probability measures and to the occupancy scheme; we start by giving precise definitions.

A non-empty set B of positive integers is called a block, and a partition of B is a denumerable family $\pi = \{\pi_1, \pi_2, \dots\}$ of pairwise disjoint sub-blocks of B such that $\cup \pi_i = B$. We write Part_B for the set of partitions of B and endow Part_B with a natural partial order : one says that a partition π is finer than another partition π' and then write $\pi \preceq \pi'$ if and only if each block π_i of π is contained into some block π'_j of π' . A sequence $(\pi(k), k \geq 0)$ of partitions is called *nested* (or, sometimes also, *refining*) if $\pi(k+1)$ is finer than $\pi(k)$ for every integer $k \geq 0$.

The occupancy scheme produces naturally random partitions. Typically, we consider some discrete probability measure $\mathbf{p} \in \text{Prob}_{\mathcal{I}}$ and the corresponding family of boxes, and we label balls by integers. Every block B can then be splitted into sub-blocks that correspond to the labels of the balls which occupy the same box. This provides a random partition of B , say $\pi_B^{\mathbf{p}}$. The latter is *exchangeable*, in the sense that its distribution is invariant under the natural action of permutations of B . Note that when B is infinite, in particular when $B = \mathbb{N}$, $\pi_B^{\mathbf{p}}$ has no singletons a.s. A fundamental theorem due to Kingman [17] (see Theorem 2.1 in [2]) claims that any exchangeable random partition of \mathbb{N} which has no singletons a.s. has the same distribution as the partition that results from some randomized version of the occupancy scheme, i.e. for which \mathbf{p} is now a random discrete probability measure. The assumption of absence of singletons can be dropped provided that one allows \mathbf{p} to be defective, i.e. to be only a sub-probability measure.

We now consider again a sequence $(\mathbf{p}(k), k \in \mathbb{N})$ of random discrete probability measures which is associated to some multiplicative cascade as in (1), and for every $k \in \mathbb{N}$, we denote by $\Pi(k)$ the random partition of \mathbb{N} induced as above by the occupancy scheme at generation k . Then $\Pi = (\Pi(k), k \geq 0)$ is a nested sequence of exchangeable random partitions of \mathbb{N} , which is Markovian. We call Π a *homogeneous fragmentation chain*. Its transition probabilities inherit the branching property from the multiplicative structure of the cascade; see Propositions 1.2 and 1.3 in [2].

For any of blocks B and B' with $B' \subseteq B$, the restriction to B' yields a natural projection $\pi \rightarrow \pi|_{B'}$ from Part_B to $\text{Part}_{B'}$. The partial order \preceq is clearly compatible with restrictions, in

the sense that $\pi \preceq \pi' \Rightarrow \pi|_{B'} \preceq \pi'|_{B'}$. Thus, if $(\pi(k), k \geq 0)$ is a nested sequence of partitions of a block B and if $B' \subseteq B$ is a smaller block, then the sequence $(\pi|_{B'}(k), k \geq 0)$ of partitions restricted to B' is again nested. A simple but nonetheless important fact is that the Markov property of a homogeneous fragmentation chain Π is preserved by restriction, in the sense that for every block $B \subseteq \mathbb{N}$, the nested sequence of partitions of B , $\Pi|_B = (\Pi|_B(k), k \geq 0)$, is still Markovian; see Lemma 3.4 in [2] for a sharper statement.

Recapitulating, the occupancy scheme enables us to associate to any random multiplicative cascade of discrete probability measures a homogeneous fragmentation chain Π . In turn, the latter provided a nested sequence of random partitions of an arbitrary block $B \subseteq \mathbb{N}$, $\Pi|_B = (\Pi|_B(k), k \geq 0)$, which is Markovian. Further, these Markov chains are coherent, in the sense that if $B' \subseteq B$, then $\Pi|_{B'} = (\Pi|_B)|_{B'}$. Roughly speaking, the statements in the preceding Section provide information about the asymptotic regimes for a homogeneous fragmentation of a finite set, when both the size n of that set and the time k at which the fragmentation process is observed tend to infinity. For example, Theorem 1 is a limit theorem for the number of components with a fixed size (like singletons, pairs, etc.), whereas Proposition 2 specifies the asymptotic behavior of the shattering time, that is the first instant at which the fragmentation process of a finite block reaches its absorbing state, i.e. the partition of that block into singletons. Finally, we also mention that, for the sake of simplicity, we have only discussed here fragmentation processes in discrete time; however our results can be shifted to homogeneous fragmentations in continuous time, using discretization techniques similar to those in developed in [3].

Appendix : Proof of the large deviations behavior

We finally proceed to the proof of Corollary 1. When f is continuous with compact support, the claim follows from Lemma 4 by approximating f with step functions. See e.g. Corollary 4 of [5] and Theorem 3 of Stone [20] for slightly stronger statements in terms of directly Riemann integrable functions with compact support. All that is needed to extend this to continuous functions with unbounded support (which fulfill the conditions of the statement) is to establish the following : If we define for some fixed $\alpha > 0$ and $\beta > \theta$

$$g_+(x) = \mathbf{1}_{\{x>0\}}x^\alpha, \quad \text{and} \quad g_-(x) = \mathbf{1}_{\{x<0\}}e^{\beta x},$$

then,

$$\sup_{k \in \mathbb{N}} \sqrt{k} e^{-\varphi(\theta)k} \sum_{i \in \mathbb{N}^k} g_\pm(kM(\theta) + \ln p_i(k)) < \infty \quad \text{a.s.} \quad (22)$$

Indeed, for every function f that fulfills the hypotheses of the statement and for every integer $\ell \geq 1$, we can find a continuous function f_ℓ with compact support such that $f_\ell \leq f \leq f_\ell + \ell^{-1}(g_+ + g_-)$, and then (22) enables us to conclude the proof by a standard argument.

Proof of (22) for g_+ : We write

$$\sum_{i \in \mathbb{N}^k} g_+(kM(\theta) + \ln p_i(k)) = \int_{\mathbb{R}_+} g_+(kM(\theta) - y) Z^{(k)}(dy) = \int_{[0, kM(\theta)]} (kM(\theta) - y)^\alpha Z^{(k)}(dy).$$

Then we pick $\theta' \in]\theta, \theta^*[$ sufficiently close to θ (as this will be explained in the sequel) and split

the last integral at $kM(\theta')$ to get

$$\int_{[0, kM(\theta')]} (kM(\theta) - y)^\alpha Z^{(k)}(dy) + \int_{]kM(\theta'), kM(\theta)]} (kM(\theta) - y)^\alpha Z^{(k)}(dy).$$

For the first integral, there is the obvious bound

$$\int_{[0, kM(\theta')]} (kM(\theta) - y)^\alpha Z^{(k)}(dy) \leq (kM(\theta))^\alpha Z^{(k)}([0, kM(\theta')]).$$

Observe from Markov inequality that $Z^{(k)}([0, a]) \leq e^{\theta a} L(\theta)^k W^{(k)}(\theta)$ for every $a > 0$, so the preceding quantity can be bounded from above by

$$(kM(\theta))^\alpha e^{\theta kM(\theta')} L(\theta)^k W^{(k)}(\theta).$$

Recall that $\varphi(\theta) = \ln L(\theta) + \theta M(\theta)$ and that $M(\theta') < M(\theta)$. As a consequence

$$(kM(\theta))^\alpha e^{\theta kM(\theta')} L(\theta)^k = o(e^{k\varphi(\theta)})/\sqrt{k}, \quad k \rightarrow \infty,$$

and since the martingale $W^{(k)}(\theta)$ remains bounded a.s., we conclude that

$$\lim_{k \rightarrow \infty} \sqrt{k} e^{-k\varphi(\theta)} \int_{[0, kM(\theta')]} (kM(\theta) - y)^\alpha Z^{(k)}(dy) = 0 \quad \text{a.s.} \quad (23)$$

For the second integral, we start from the bound

$$\begin{aligned} & \int_{]kM(\theta'), kM(\theta)]} (kM(\theta) - y)^\alpha Z^{(k)}(dy) \\ & \leq \sum_{0 \leq j \leq k(M(\theta) - M(\theta'))} j^\alpha Z^{(k)}([kM(\theta) - j, kM(\theta) - j + 1]). \end{aligned}$$

For indices $0 \leq j \leq k(M(\theta) - M(\theta'))$, we define

$$\theta_{k,j} = M^{-1}(M(\theta) - j/k) \in [\theta, \theta'], \quad (24)$$

so that

$$kM(\theta_{k,j}) = kM(\theta) - j.$$

We observe that

$$\begin{aligned} Z^{(k)}([kM(\theta) - j, kM(\theta) - j + 1]) &= L(\theta_{k,j})^k \int_{kM(\theta) - j}^{kM(\theta) - j + 1} e^{\theta_{k,j} y} Z_{\theta_{k,j}}^{(k)}(dy) \\ &\leq e^{\theta_{k,j}(kM(\theta) - j + 1)} L(\theta_{k,j})^k Z_{\theta_{k,j}}^{(k)}([kM(\theta) - j, kM(\theta) - j + 1]) \\ &\leq e^{(1-j)\theta} \left(e^{\theta_{k,j} M(\theta)} L(\theta_{k,j}) \right)^k Z_{\theta_{k,j}}^{(k)}([kM(\theta_{k,j}), kM(\theta_{k,j}) + 1]), \end{aligned}$$

where for the last inequality, we use the fact that $\theta_{k,j} \geq \theta$.

Recall that $\ln L$ is convex and has derivative $-M$. Since $\theta \leq \theta_{k,j} \leq \theta'$, we have

$$\ln L(\theta_{k,j}) \leq \ln L(\theta) - M(\theta')(\theta_{k,j} - \theta),$$

and since $\varphi(\theta) = \ln L(\theta) + \theta M(\theta)$, this yields

$$\theta_{k,j} M(\theta) + \ln L(\theta_{k,j}) \leq \varphi(\theta) + (\theta_{k,j} - \theta)(M(\theta) - M(\theta')).$$

As the mean function M is locally a regular diffeomorphism, we see from (24) that there is some finite constant C (which is independent of k and j) such that $\theta_{k,j} - \theta \leq Cj/k$. Further, we also have that $M(\theta) - M(\theta') \leq \theta/2C$ provided that we choose θ' sufficiently close to θ . Then

$$\left(e^{\theta_{k,j} M(\theta)} L(\theta_{k,j}) \right)^k \leq e^{j\theta/2} \exp(k\varphi(\theta)),$$

and thus

$$\sum_{0 \leq j \leq k(M(\theta) - M(\theta'))} j^\alpha e^{(1-j)\theta} \left(e^{\theta_{k,j} M(\theta)} L(\theta_{k,j}) \right)^k = O(\exp(k\varphi(\theta))).$$

On the other hand, we know from Lemma 4 that there exists an a.s. finite random variable ξ such that

$$Z_{\theta_{k,j}}^{(k)}([kM(\theta_{k,j}), kM(\theta_{k,j}) + 1]) \leq k^{-1/2} \xi.$$

Putting the pieces together, we conclude that

$$\sup_{k \in \mathbb{N}} \sqrt{k} e^{-k\varphi(\theta)} \int_{[kM(\theta'), kM(\theta)]} (kM(\theta) - y)^\alpha Z^{(k)}(dy) < \infty \quad \text{a.s.}$$

Combining with (23), we have thus checked that (22) does hold for g_+ . \square

The proof of the bound (22) for g_- follows a similar route; however it may be useful to spell out the main steps.

Proof of (22) for g_- : We start with

$$\sum_{i \in \mathbb{N}^k} g_-(kM(\theta) + \ln p_i(k)) = \int_{[0, \infty[} e^{-\beta y} Z^{(k)}(kM(\theta) + dy).$$

We pick $\theta' \in]\theta_*, \theta[$ sufficiently close to θ and split this integral at $k(M(\theta') - M(\theta))$.

For indices $0 \leq j < k(M(\theta') - M(\theta))$, we introduce

$$\theta_{k,j} = M^{-1}(M(\theta) + j/k) \in [\theta', \theta], \quad (25)$$

so that

$$kM(\theta_{k,j}) = kM(\theta) + j.$$

We observe that

$$\begin{aligned} & \int_{[j, j+1[} e^{-\beta y} Z^{(k)}(kM(\theta) + dy) \\ &= L(\theta_{k,j})^k \int_{[j, j+1[} e^{-\beta y} e^{\theta_{k,j} kM(\theta) + y} Z_{\theta_{k,j}}^{(k)}(kM(\theta) + dy) \\ &\leq L(\theta_{k,j})^k e^{-j(\beta - \theta_{k,j})} e^{k\theta_{k,j} M(\theta)} Z_{\theta_{k,j}}^{(k)}([kM(\theta) + j, kM(\theta) + j + 1]) \\ &\leq e^{-j(\beta - \theta)} \left(e^{\theta_{k,j} M(\theta)} L(\theta_{k,j}) \right)^k Z_{\theta_{k,j}}^{(k)}([kM(\theta_{k,j}), kM(\theta_{k,j}) + 1]), \end{aligned}$$

where for the last inequality, we use the fact that $\theta_{k,j} \leq \theta$.

Because $\ln L$ is convex, has derivative $-M$ and $\theta' \leq \theta_{k,j} \leq \theta$, there is the inequality

$$\ln L(\theta_{k,j}) \leq \ln L(\theta) + M(\theta')(\theta - \theta_{k,j}),$$

which yields

$$\theta_{k,j}M(\theta) + \ln L(\theta_{k,j}) \leq \varphi(\theta) + (\theta - \theta_{k,j})(M(\theta') - M(\theta)).$$

We see from (25) that there is some finite constant C (which is independent of k and j) such that $\theta - \theta_{k,j} \leq Cj/k$, and that $M(\theta') - M(\theta) \leq (\beta - \theta)/2C$ provided that we choose θ' sufficiently close to θ . Then

$$\left(e^{\theta_{k,j}M(\theta)} L(\theta_{k,j}) \right)^k \leq e^{j(\beta - \theta)/2} \exp(k\varphi(\theta)),$$

and thus

$$\sum_{0 \leq j < k(M(\theta') - M(\theta))} e^{-j(\beta - \theta)} \left(e^{\theta_{k,j}M(\theta)} L(\theta_{k,j}) \right)^k = O(\exp(k\varphi(\theta))).$$

Further Lemma 4 ensures the existence of an a.s. finite random variable ξ such that

$$Z_{\theta_{k,j}}^{(k)}([kM(\theta_{k,j}), kM(\theta_{k,j}) + 1]) \leq k^{-1/2}\xi,$$

and then we can conclude that

$$\sup_{k \in \mathbb{N}} \sqrt{k} e^{-k\varphi(\theta)} \int_{[0, k(M(\theta') - M(\theta))]} e^{-\beta y} Z^{(k)}(kM(\theta) + dy) < \infty \quad \text{a.s.} \quad (26)$$

For the remaining integral, we note that

$$\begin{aligned} & \int_{[k(M(\theta') - M(\theta)), \infty[} e^{-\beta y} Z^{(k)}(kM(\theta) + dy) \\ & \leq e^{-(\beta - \theta)k(M(\theta') - M(\theta))} \int_{[k(M(\theta') - M(\theta)), \infty[} e^{-\theta y} Z^{(k)}(kM(\theta) + dy) \\ & \leq e^{-(\beta - \theta)k(M(\theta') - M(\theta))} e^{k\theta M(\theta)} L(\theta)^k W^{(k)}(\theta). \end{aligned}$$

We readily deduce that

$$\lim_{k \rightarrow \infty} \sqrt{k} e^{-k\varphi(\theta)} \int_{[k(M(\theta') - M(\theta)), \infty[} e^{-\beta y} Z^{(k)}(kM(\theta) + dy) = 0 \quad \text{a.s.},$$

and combining with (26), this establishes that (22) holds for g_- . □

Acknowledgment. I would like to thank an anonymous referee for having carefully checked of the first draft of this work, and especially for pointing at some errors.

References

- [1] R. R. Bahadur (1960). On the number of distinct values in a large sample from an infinite discrete distribution. *Proc. Nat. Inst. Sci. India Part A* **26**, 67-75.
- [2] J. Bertoin (2006). *Random Fragmentation and Coagulation Processes*. Cambridge University Press, Cambridge.
- [3] J. Bertoin and A. Rouault (2005). Discretization methods for homogeneous fragmentations. *J. London Math. Soc.* **72**, 91-109.
- [4] J. D. Biggins (1977). Martingale convergence in the branching random walk. *J. Appl. Probability* **14**, 25-37.
- [5] J. D. Biggins (1992). Uniform convergence of martingales in the branching random walk. *Ann. Probab.* **20**, 137-151.
- [6] S. Boucheron and D. Gardy (1997). An urn model from learning theory. *Random Struct. Algorithms* **10**, 43-69.
- [7] J. Bunge and M. Fitzpatrick (1993). Estimating the number of species : a review. *J. Am. Statist. Assoc.* **88**, 364-373.
- [8] M. Dutko (1989). Central limit theorems for infinite urn models. *Ann. Probab.* **17**, 1255-1263.
- [9] D. Gardy (2002). Occupancy urn models in the analysis of algorithms. *J. Stat. Plann. Inference* **101**, 95-105.
- [10] A. Gnedin, B. Hansen and J. Pitman (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys* **4**, 146-171.
- [11] A. Gnedin, J. Pitman and M. Yor (2006). Asymptotic laws for compositions derived from transformed subordinators. *Ann. Probab.* **34**, 468-492.
- [12] F. He and K. J. Gaston (2000). Estimating species abundance from occurrence. *The American Naturalist* **156-5**, 553-559.
- [13] L. Holst (1986). On birthday, collectors', occupancy and other classical urn problems. *Int. Stat. Rev.* **54**, 15-27.
- [14] H.-K. Hwang and S. Janson (2007+). Local limit theorems for finite and infinite urn models. To appear in *Ann. Probab.*
- [15] N. L. Johnson and S. Kotz (1977). *Urn Models and Their Application. An Approach to Modern Discrete Probability Theory*. John Wiley & Sons, New York-London-Sydney.
- [16] S. Karlin (1967). Central limit theorems for certain infinite urn schemes. *J. Math. Mech.* **17**, 373-401.

- [17] J. F. C. Kingman(1982). The coalescent. *Stochastic Process. Appl.* **13**, 235-248.
- [18] V. F. Kolchin, B. A. Sevast'yanov and V. P. Chistyakov (1978). *Random Allocations*. John Wiley & Sons, New York-London-Sydney.
- [19] Q. S. Liu (2000). On generalized multiplicative cascades. *Stochastic Process. Appl.* **86**, 263-286.
- [20] C. Stone (1967). On local and ratio limit theorems. *Proc. 5th Berkeley Sympos. math. Statist. Probab.* **2**, 217-224.